Cox process functional learning

Gérard Biau, Benoît Cadre, Quentin Paris

Cox Process Learning

Gérard Biau
Université Pierre et Marie Curie[1] *& Ecole Normale Supérieure*[2]*, France*
gerard.biau@upmc.fr

Benoît Cadre
IRMAR, ENS Cachan Bretagne, CNRS, UEB, France[3]
cadre@bretagne.ens-cachan.fr

Quentin Paris
IRMAR, ENS Cachan Bretagne, CNRS, UEB, France
quentin.paris@bretagne.ens-cachan.fr

Abstract

This article addresses the problem of supervised classification of Cox process trajectories, whose random intensity is driven by some exogenous random covariable. The classification task is achieved through a regularized convex empirical risk minimization procedure, and a nonasymptotic oracle inequality is derived. We show that the algorithm provides a Bayes-risk consistent classifier. Furthermore, it is proved that the classifier converges at a rate which adapts to the unknown regularity of the intensity process. Our results are obtained by taking advantage of martingale and stochastic calculus arguments, which are natural in this context and fully exploit the functional nature of the problem.

Index Terms — Cox process, supervised classification, oracle inequality, consistency, regularization, stochastic calculus.

2010 Mathematics Subject Classification: 62G05, 62G20.

1 Introduction

1.1 The model

This article deals with the supervised classification problem of Cox process trajectories. Let us first recall that a random process $N = (N_t)_{t\in[0,T]}$ $(T >$

[1]Research partially supported by the French National Research Agency (grant ANR-09-BLAN-0051-02 "CLARA") and by the Institut universitaire de France.

[2]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

[3]Research sponsored by the French National Research Agency (grant ANR-09-BLAN-0051-02 "CLARA").

0) is a counting process if its trajectories are, with probability one, right-continuous and piecewise constant, if it starts at 0, and if the jump size of N at time t is, with probability one, either 0 or 1. In more pedestrian terms, the process N starts at $N_0 = 0$ and stays at level 0 until some random time T_1 when it jumps to $N_{T_1} = 1$. It then stays at level 1 until another random time T_2 when it jumps to the value $N_{T_2} = 2$, and so on. The random times $T_1, T_2, \ldots$ are referred to as the jump times of N. In point process theory and its applications, such as for example in neuronal potential generation (Krumin and Shoham, 2009) and credit risk theory (Lando, 1998), an important role is played by a particular class of counting processes known as Cox processes (or doubly stochastic Poisson processes). In a word, we say that N is a Cox process with (random) intensity process $\lambda = (\lambda_t)_{t\in[0,T]}$ if λ is predictable (roughly, this means that λ is a left-continuous and adapted process), if λ has summable trajectories, and if the conditional distribution of N given λ is that of a Poisson process with intensity function λ. Thus, a Cox process is a generalization of a Poisson process where the time-dependent intensity λ is itself a stochastic process. The process is named after the statistician David Cox, who first published the model in 1955 (Cox, 1955). For details, the reader is referred to Grimmett and Stirzaker (2001); Jacod and Shiryaev (2003); Durrett (2010) or any other standard textbook on the subject.

Going back to our supervised classification problem, we consider a prototype random triplet (X, Z, Y), where Y is a binary label taking the values ± 1 with respective positive probabilities p_+ and p_- ($p_+ + p_- = 1$). In this model, $Z = (Z_t)_{t\in[0,T]}$ plays the role of a d-dimensional random covariable (process), whereas $X = (X_t)_{t\in[0,T]}$ is a mixture of two Cox processes. More specifically, it is assumed that Z is independent of Y and that, conditionally on $Y = 1$ (resp., $Y = -1$), X is a Cox process with intensity $(\lambda_+(t, Z_t))_{t\in[0,T]}$ (resp., $(\lambda_-(t, Z_t))_{t\in[0,T]}$). A typical example is when the hazard rate processes are expressed as

$$\lambda_+(t, Z_t) = \lambda_1(t) \exp\left(\theta_1^T Z_t\right) \quad \text{and} \quad \lambda_-(t, Z_t) = \lambda_2(t) \exp\left(\theta_2^T Z_t\right)$$

(vectors are in column format and T denotes a transpose). In this model, θ_1 and θ_2 are unknown regression coefficients, and $\lambda_1(t)$ and $\lambda_2(t)$, the underlying baseline hazards, are unknown and unspecified nonnegative functions. This is the celebrated Cox (1972) proportional hazard model, which is widely used in modern survival analysis.

However, in the sequel, apart from some minimal smoothness assumptions, we do not impose any particular parametric or semi-parametric model for the functions λ_+ and λ_-. On the other hand, it will be assumed that the

observation of the trajectories of X is stopped after its u-th jump, where u is some known, prespecified, positive integer. Thus, formally, we are to replace X and Z by X^τ and Z^τ, where $\tau = \inf\{t \in [0,T] : X_t = u\}$ (stopping time), $X_t^\tau = X_{t\wedge\tau}$ and $Z_t^\tau = Z_{t\wedge\tau}$. (Notation $t_1 \wedge t_2$ means the minimum of t_1 and t_2 and, by convention, $\inf\emptyset = 0$.) Stopping the observation of X after its u-th jump is essentially a technical requirement, with no practical incidence insofar u may be chosen arbitrarily large. However, it should be stressed that with this assumption, X^τ is, with probability one, nicely bounded from above by u. Additionally, to keep things simple, we suppose that each Z_t takes its values in $[0,1]^d$.

Our objective is to learn the relation between (X^τ, Z^τ) and Y within the framework of supervised classification (see, e.g., Devroye et al., 1996). To this aim, denote by $\mathcal{X}$ (resp., $\mathcal{Z}$) the space of real-valued, positive, nondecreasing, piecewise constant, and right-continuous functions on $[0,T]$ with jumps of size 1 (resp., the state space of Z). Given a training dataset of n i.i.d. observation/label pairs $\mathcal{D}_n = (X_1^{\tau_1}, Z_1^{\tau_1}, Y_1), \ldots, (X_n^{\tau_n}, Z_n^{\tau_n}, Y_n)$, distributed as (and independent of) the prototype triplet (X^τ, Z^τ, Y), the problem is to design a decision rule $g_n : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$, based on $\mathcal{D}_n$, whose role is to assign a label to each possible new instance of the observation (X^τ, Z^τ). The classification strategy that we propose is based on empirical convex risk minimization. It is described in the next subsection.

We have had several motivations for undertaking this study. First of all, much effort has been spent in recent years in deriving models for functional data analysis, a branch of statistics that analyzes data providing information about curves, surfaces or anything else varying over a continuum (the monograph by Ramsay and Silverman, 2005, offers a comprehensive introduction to the domain). Curiously, despite a huge research activity in this area, few attempts have been made to connect the rich theory of stochastic processes with functional data analysis (interesting references towards this direction are Bouzas et al., 2006; Illian et al., 2006; Shuang et al., 2013; Zhu et al., 2011; Cadre, 2012; Denis, 2012). We found that the martingale and stochastic calculus theory—which emerges naturally from the formulation of our classification problem—could be used very efficiently to give proofs whose basic ideas are simple and may serve as a starting point for more dialogue among these parties. Secondly, we found it useful to know how the modern aspects of statistical learning theory could be adapted to the context of point processes. With this respect, our point of view differs from more classical approaches, which are mainly devoted to the statistical inference in Cox (1972) proportional hazard model (e.g., Cox, 1975; Andersen and Gill, 1982; O'Sullivan, 1993, and the references therein). Finally, even if our approach

is of theoretical nature and we have no data to show, we believe that there are a number of practical examples where our model and procedure could be used. Just think, for example, to cancer patients regularly admitted to hospital (process X_t), followed by their personal data files (process Z_t), and for which doctors want to make a diagnostic (-1=aggravation, $+1$=remission, for example).

1.2 Classification strategy

In order to describe our classification procedure, some more notation is required. The performance of a classifier $g_n : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$ is measured by the probability of error

$$L(g_n) = \mathbb{P}\left(g_n(X^\tau, Z^\tau) \neq Y \,|\, \mathcal{D}_n\right),$$

and the minimal possible probability of error is the Bayes risk, denoted by

$$L^\star = \inf_g L(g) = \mathbb{E}\min\left[\eta(X^\tau, Z^\tau), 1 - \eta(X^\tau, Z^\tau)\right].$$

In the identity above, the infimum is taken over all measurable classifiers $g : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$, and $\eta(X^\tau, Z^\tau) = \mathbb{P}(Y = 1 \,|\, X^\tau, Z^\tau)$ denotes the posterior probability function. The infimum is achieved by the Bayes classifier

$$g^\star(X^\tau, Z^\tau) = \mathrm{sign}(2\eta(X^\tau, Z^\tau) - 1),$$

where $\mathrm{sign}(t) = 1$ for $t > 0$ and -1 otherwise. Our first result (Theorem 2.1) shows that

$$\eta(X^\tau, Z^\tau) = \frac{p_+}{p_- e^{-\xi} + p_+},$$

where ξ is the $\sigma(X^\tau, Z^\tau)$-measurable random variable defined by

$$\xi = \int_0^{T\wedge\tau} (\lambda_- - \lambda_+)(s, Z_s)\mathrm{d}s + \int_0^{T\wedge\tau} \ln\frac{\lambda_+}{\lambda_-}(s, Z_s)\mathrm{d}X_s$$

(For all $x \in \mathcal{X}$ and any function g, the notation $\int g(s)dx_s$ refers to the integral of g with respect to the Stieltjes measure associated with the nondecreasing function x. We refer the reader to Chapter 0 in Revuz and Yor, 2005, for more details.) An important consequence is that the Bayes rule associated with our decision problem takes the simple form

$$g^\star(X^\tau, Z^\tau) = \mathrm{sign}\left(\xi - \ln\frac{p_-}{p_+}\right).$$

Next, let $(\varphi_j)_{j\geq 1}$ be a countable dictionary of measurable functions defined on $[0,T]\times[0,1]^d$. Assuming that both $\lambda_- - \lambda_+$ and $\ln\frac{\lambda_+}{\lambda_-}$ belong to the span of the dictionary, we see that

$$\xi = \sum_{j\geq 1}\left[a_j^\star \int_0^{T\wedge\tau} \varphi_j(s,Z_s)\mathrm{d}s + b_j^\star \int_0^{T\wedge\tau} \varphi_j(s,Z_s)\mathrm{d}X_s\right],$$

where $(a_j^\star)_{j\geq 1}$ and $(b_j^\star)_{j\geq 1}$ are two sequences of unknown real coefficients. Thus, for each positive integer B, it is quite natural to introduce the class $\mathcal{F}_B$ of real-valued functions $f:\mathcal{X}\times\mathcal{Z}\to\mathbb{R}$, defined by

$$\mathcal{F}_B = \left\{ f = \sum_{j=1}^B [a_j\Phi_j + b_j\Psi_j] + c : \max\left(\sum_{j=1}^B |a_j|, \sum_{j=1}^B |b_j|, |c|\right) \leq B \right\},$$

where

$$\Phi_j(x,z) = \int_0^{T\wedge\tau(x)} \varphi_j(s,z_s)\mathrm{d}s, \qquad \Psi_j(x,z) = \int_0^{T\wedge\tau(x)} \varphi_j(s,z_s)\mathrm{d}x_s,$$

and, by definition, $\tau(x) = \inf\{t\in[0,T] : x_t = u\}$.

Each $f\in\mathcal{F}_B$ defines a classifier g_f by $g_f = \mathrm{sign}(f)$. To simplify notation, we write $L(f) = L(g_f) = \mathbb{P}(g_f(X^\tau, Z^\tau)\neq Y)$, and note that

$$\mathbb{E}\mathbf{1}_{[-Yf(X^\tau,Z^\tau)>0]} \leq L(f) \leq \mathbb{E}\mathbf{1}_{[-Yf(X^\tau,Z^\tau)\geq 0]}.$$

Therefore, the minimization of the probability of error $L(f)$ over $f\in\mathcal{F}_B$ is approximately equivalent to the minimization of the expected 0-1 loss $\mathbf{1}_{[.\geq 0]}$ of $-Yf(X^\tau,Z^\tau)$. The parameter B may be regarded as an $\mathbb{L}^1$-type smoothing parameter. Large values of B improve the approximation properties of the class $\mathcal{F}_B$ at the price of making the estimation problem more difficult. Now, given the sample $\mathcal{D}_n$, it is reasonable to consider an estimation procedure based on minimizing the sample mean

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{[-Y_i f(X_i^{\tau_i}, Z_i^{\tau_i})\geq 0]},$$

of the 0-1 loss.

It is now well established, however, that such a procedure is computationally intractable as soon as the class $\mathcal{F}_B$ is nontrivial, since the 0-1 loss function $\mathbf{1}_{[.\geq 0]}$ is nonconvex. A genuine attempt to circumvent this difficulty is to base the minimization procedure on a convex surrogate ϕ of the loss $\mathbf{1}_{[.\geq 0]}$.

Such convexity-based methods, inspired by the pioneering works on boosting (Freund, 1995; Schapire, 1990; Freund and Schapire, 1997), have now largely displaced earlier nonconvex approaches in the machine learning literature (see, e.g., Blanchard et al., 2003; Lugosi and Vayatis, 2004; Zhang, 2004; Bartlett et al., 2006, and the references therein).

It turns out that in our Cox process context, the choice of the logit surrogate loss $\phi(t) = \ln_2(1+e^t)$ is the most natural one. This will be clarified in Section 2 by connecting the empirical risk minimization procedure and the maximum likelihood principle. Thus, with this choice, the corresponding risk functional and empirical risk functional are defined by

$$A(f) = \mathbb{E}\phi\left(-Yf(X^\tau, Z^\tau)\right) \quad \text{and} \quad A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(-Y_i f(X_i^{\tau_i}, Z_i^{\tau_i})\right).$$

Given a nondecreasing sequence $(B_k)_{k\geq 1}$ of integer-valued smoothing parameters, the primal estimates we consider take the form

$$\hat{f}_k \in \operatorname*{arg\,min}_{f\in\mathcal{F}_{B_k}} A_n(f).$$

(Note that the minimum may not be achieved in $\mathcal{F}_{B_k}$. However, to simplify the arguments, we implicitly assume that the minimum indeed exists. All proofs may be adjusted, in a straightforward way, to handle approximate minimizers of the empirical risk functional). Starting from the collection $(\hat{f}_k)_{k\geq 1}$, the final estimate uses a value of k chosen empirically, by minimizing a penalized version of the empirical risk $A_n(\hat{f}_k)$. To achieve this goal, consider a penalty (or regularization) function $\mathrm{pen} : \mathbb{N}^\star \to \mathbb{R}_+$ to be precised later on. Then the resulting penalized estimate $\hat{f}_n = \hat{f}_{\hat{k}}$ has

$$\hat{k} \in \operatorname*{arg\,min}_{k\geq 1} \left[A_n(\hat{f}_k) + \mathrm{pen}(k)\right].$$

The role of the penalty is to compensate for overfitting and helps finding an adequate value of k. For larger values of k, the class $\mathcal{F}_{B_k}$ is larger, and therefore $\mathrm{pen}(k)$ should be larger as well.

By a careful choice of the regularization term, specified in Theorem 2.2, one may find a close-to-optimal balance between estimation and approximation errors and investigate the probability of error $L(\hat{f}_n)$ of the classifier $g_{\hat{f}_n}$ induced by the penalized estimate. Our conclusion asserts that $\hat{f}_n$ adapts nicely to the unknown smoothness of the problem, in the sense that with probability at least $1 - 1/n^2$,

$$L(\hat{f}_n) - L^\star = \mathrm{O}\left(\frac{\ln n}{n}\right)^{\frac{\beta}{2\beta+16}},$$

where β is some Sobolev-type regularity measure pertaining to λ_+ and λ_-. For the sake of clarity, proofs are postponed to Section 3. An appendix at the end of the paper recalls some important results by Blanchard et al. (2008) and Koltchinskii (2011) on model selection and suprema of Rademacher processes, together with more technical stochastic calculus material.

2 Results

As outlined in the introduction, our first result shows that the posterior probabilities $\mathbb{P}(Y = \pm 1 | X^\tau, Z^\tau)$ have a simple form. The crucial result that is needed here is Lemma A.1 which uses martingale and stochastic calculus arguments. For more clarity, this lemma has been postponed to the Appendix section. Recall that both p_+ and p_- are (strictly) positive and satisfy $p_+ + p_- = 1$.

Theorem 2.1 *Let ξ be the $\sigma(X^\tau, Z^\tau)$-measurable random variable defined by*

$$\xi = \int_0^{T\wedge\tau} (\lambda_- - \lambda_+)(s, Z_s)\mathrm{d}s + \int_0^{T\wedge\tau} \ln\frac{\lambda_+}{\lambda_-}(s, Z_s)\mathrm{d}X_s.$$

Then

$$\mathbb{P}(Y = 1 | X^\tau, Z^\tau) = \frac{p_+}{p_- e^{-\xi} + p_+} \quad \textit{and} \quad \mathbb{P}(Y = -1 | X^\tau, Z^\tau) = \frac{p_-}{p_+ e^{\xi} + p_-}.$$

This result, which is interesting by itself, sheds an interesting light on the Cox process classification problem. To see this, fix $Y_1 = y_1, \ldots, Y_n = y_n$, and observe that the conditional likelihood of the model is

$$\begin{aligned}\mathcal{L}_n &= \prod_{i=1}^n \mathbb{P}(Y_i = y_i | X_i^{\tau_i}, Z_i^{\tau_i}) \\ &= \prod_{i=1}^n \left(\frac{p_+}{p_- e^{-y_i\xi_i} + p_+}\right)^{\mathbf{1}_{[y_i=1]}} \left(\frac{p_-}{p_+ e^{-y_i\xi_i} + p_-}\right)^{\mathbf{1}_{[y_i=-1]}},\end{aligned}$$

where of course

$$\xi_i = \int_0^{T\wedge\tau} (\lambda_- - \lambda_+)(s, Z_{i,s})\mathrm{d}s + \int_0^{T\wedge\tau} \ln\frac{\lambda_+}{\lambda_-}(s, Z_{i,s})\mathrm{d}X_{i,s}.$$

Therefore, the log-likelihood takes the form

$$\begin{aligned}
\ln \mathcal{L}_n &= \sum_{i=1}^{n} \left[\ln \left(\frac{p_+}{p_- e^{-y_i \xi_i} + p_+} \right) \mathbf{1}_{[y_i=1]} + \ln \left(\frac{p_-}{p_+ e^{-y_i \xi_i} + p_-} \right) \mathbf{1}_{[y_i=-1]} \right] \\
&= -\sum_{i=1}^{n} \left[\ln \left(1 + \frac{p_-}{p_+} e^{-y_i \xi_i} \right) \mathbf{1}_{[y_i=1]} + \ln \left(1 + \frac{p_+}{p_-} e^{-y_i \xi_i} \right) \mathbf{1}_{[y_i=-1]} \right] \\
&= -\sum_{i=1}^{n} \ln \left(1 + \left(\frac{p_-}{p_+} \right)^{y_i} e^{-y_i \xi_i} \right) \\
&= -\sum_{i=1}^{n} \ln \left(1 + \exp \left[-y_i \left(\xi_i - \ln \frac{p_-}{p_+} \right) \right] \right).
\end{aligned}$$

Thus, letting $\phi(t) = \ln_2(1 + e^t)$, we obtain

$$\ln \mathcal{L}_n = -\ln 2 \sum_{i=1}^{n} \phi \left(-y_i \left(\xi_i - \ln \frac{p_-}{p_+} \right) \right). \tag{2.1}$$

Since the ξ_i's, p_+ and p_- are unknown, the natural idea, already alluded to in the introduction, is to expand $\lambda_- - \lambda_+$ and $\ln \frac{\lambda_+}{\lambda_-}$ on the dictionary $(\varphi_j)_{j \geq 1}$. To this end, we introduce the class $\mathcal{F}_B$ of real-valued functions

$$\mathcal{F}_B = \left\{ f = \sum_{j=1}^{B} [a_j \Phi_j + b_j \Psi_j] + c : \max \left(\sum_{j=1}^{B} |a_j|, \sum_{j=1}^{B} |b_j|, |c| \right) \leq B \right\},$$

where B is a positive integer,

$$\Phi_j(x, z) = \int_0^{T \wedge \tau(x)} \varphi_j(s, z_s) \mathrm{d}s, \quad \text{and} \quad \Psi_j(x, z) = \int_0^{T \wedge \tau(x)} \varphi_j(s, z_s) \mathrm{d}x_s.$$

For a nondecreasing sequence $(B_k)_{k \geq 1}$ of integer-valued smoothing parameters and for each $k \geq 1$, we finally select $\hat{f}_k \in \mathcal{F}_{B_k}$ for which the log-likelihood (2.1) is maximal. Clearly, such a maximization strategy is strictly equivalent to minimizing over $f \in \mathcal{F}_{B_k}$ the empirical risk

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} \phi \left(-Y_i f(X_i^{\tau_i}, Z_i^{\tau_i}) \right).$$

This remark reveals the deep connection between our Cox process learning model and the maximum likelihood principle. In turn, it justifies the logit loss $\phi(t) = \ln_2(1 + e^t)$ as the natural surrogate candidate to the nonconvex

0-1 classification loss. (Note that the $\ln 2$ term is introduced for technical reasons only and plays no role in the analysis).

As for now, denoting by $\|.\|_\infty$ the functional supremum norm, we assume that there exists a positive constant L such that, for each $j \geq 1$, $\|\varphi_j\|_\infty \leq L$. It immediately follows that for all integers $B \geq 1$, the class $\mathcal{F}_B$ is uniformly bounded by UB, where $U = 1 + (T + u)L$. We are now ready to state our main theorem, which offers a bound on the difference $A(\hat{f}_n) - A(f^\star)$.

Theorem 2.2 *Let $(B_k)_{k\geq 1}$ be a nondecreasing sequence of positive integers such that $\sum_{k\geq 1} B_k^{-\alpha} \leq 1$ for some $\alpha > 0$. For all $k \geq 1$, let*

$$R_k = A_k^2 B_k C_k + \frac{\sqrt{A_k}}{C_k},$$

where

$$A_k = UB_k\phi'(UB_k) \quad \text{and} \quad C_k = 2(\phi(UB_k) + 1 - \ln 2).$$

Then there exists a universal constant $C > 0$ such that if the penalty pen : $\mathbb{N}^\star \to \mathbb{R}_+$ *satisfies*

$$\text{pen}(k) \geq C\left[R_k \frac{\ln n}{n} + \frac{C_k(\alpha \ln B_k + \delta + \ln 2)}{n}\right]$$

for some $\delta > 0$, one has, with probability at least $1 - e^{-\delta}$,

$$A(\hat{f}_n) - A(f^\star) \leq 2 \inf_{k\geq 1}\left\{\inf_{f\in\mathcal{F}_{B_k}} (A(f) - A(f^\star)) + \text{pen}(k)\right\}. \tag{2.2}$$

Some remarks are in order. At first, we note that Theorem 2.2 provides us with an oracle inequality which shows that, for each B_k, the penalized estimate does almost as well as the best possible classifier in the class $\mathcal{F}_{B_k}$, up to a term of the order $\ln n/n$. It is stressed that this remainder term tends to 0 at a much faster rate than the standard $(1/\sqrt{n})$-term suggested by a standard uniform convergence argument (see, e.g., Lugosi and Vayatis, 2004). This is a regularization effect which is due to the convex loss ϕ. In fact, proof of Theorem 2.2 relies on the powerful model selection machinery presented in Blanchard et al. (2008) coupled with modern empirical process theory arguments developed in Koltchinskii (2011). We also emphasize that a concrete but suboptimal value of the constant C may be deduced from the proof, but that no attempt has been made to optimize this constant. Next, observing that, for the logit loss,

$$\phi'(t) = \frac{1}{\ln 2(e^{-t} + 1)},$$

we notice that a penalty behaving as B_k^4 is sufficient for the oracle inequality of Theorem 2.2 to hold. This corresponds to a regularization function proportional to the fourth power of the $\mathbb{L}^1$-norm of the collection of coefficients defining the base class functions. Such regularizations have been explored by a number of authors in recent years, specifically in the context of sparsity and variable selection (see, e.g., Tibshirani, 1996; Candès and Tao, 2005; Bunea et al., 2007; Bickel et al., 2009). With this respect, our approach is close to the view of Massart and Meynet (2011), who provide information about the Lasso as an $\mathbb{L}^1$-regularization procedure per se, together with sharp $\mathbb{L}^1$-oracle inequalities. Let us finally mention that the result of Theorem 2.2 can be generalized, with more technicalities, to other convex loss functions by following, for example, the arguments presented in Bartlett et al. (2006).

If we are able to control the approximation term $\inf_{f\in\mathcal{F}_{B_k}}(A(f)-A(f^\star))$ in inequality (2.2), then it is possible to give an explicit rate of convergence to 0 for the quantity $A(\hat{f}_n)-A(f^\star)$. This can be easily achieved by assuming, for example, that $(\varphi_j)_{j\geq 1}$ is an orthonormal basis and that both combinations $\lambda_- - \lambda_+$ and $\ln\frac{\lambda_+}{\lambda_-}$ enjoy some Sobolev-type regularity with respect to this basis. Also, the following additional assumption will be needed:

Assumption A. There exists a measure μ on $[0,1]^d$ and a constant $D>0$ such that, for all $t\in[0,T]$, the distribution of Z_t has a density with respect to μ which is uniformly bounded by D. In addition, λ_- and λ_+ are both $[\varepsilon, D]$-valued for some $\varepsilon>0$.

Proposition 2.1 *Assume that Assumption* **A** *holds. Assume, in addition, that $(\varphi_j)_{j\geq 1}$ is an orthonormal basis of $\mathbb{L}^2(\mathrm{d}s\otimes\mu)$, where $\mathrm{d}s$ stands for the Lebesgue measure on $[0,T]$, and that both $\lambda_- - \lambda_+$ and $\ln\frac{\lambda_+}{\lambda_-}$ belong to the ellipsoïd*

$$\mathcal{W}(\beta,M)=\left\{f=\sum_{j=1}^{\infty}a_j\varphi_j : \sum_{j=1}^{\infty}j^{2\beta}a_j^2\leq M^2\right\},$$

for some fixed $\beta\in\mathbb{N}^\star$ and $M>0$. Then, letting

$$\lambda_- - \lambda_+ = \sum_{j=1}^{\infty}a_j^\star\varphi_j \quad \textit{and} \quad \ln\frac{\lambda_+}{\lambda_-}=\sum_{j=1}^{\infty}b_j^\star\varphi_j,$$

we have, for all $B\geq\max(M^2,\ln\frac{p_+}{p_-})$,

$$\inf_{f\in\mathcal{F}_B}(A(f)-A(f^\star))\leq 2D\sqrt{\frac{M\|a^\star\|_2}{T\mu\left([0,1]^d\right)B^\beta}}+4D\sqrt{\frac{M\|b^\star\|_2}{B^\beta}},$$

where $\|a^\star\|_2^2=\sum_{j=1}^{\infty}a_j^{\star 2}$ and $\|b^\star\|_2^2=\sum_{j=1}^{\infty}b_j^{\star 2}$.

A careful inspection of Theorem 2.2 and Proposition 2.1 reveals that for the choice $B_k = \lfloor k^{2/\alpha} \rfloor$, $\delta = 2 \ln n$, and some constant $C' > 0$ depending on T, u, L, α, M, μ, D, $a^\star$ and $b^\star$ we have, provided n is large enough,

$$A(\hat{f}_n) - A(f^\star) \leq C' \left(\frac{\ln n}{n} \right)^{\frac{\beta}{\beta+8}},$$

with probability at least $1 - 1/n^2$.

Of course, our main concern is not the behavior of the expected risk $A(\hat{f}_n)$ but the probability of error $L(\hat{f}_n)$ of the corresponding classifier. Fortunately, the difference $L(\hat{f}_n) - L^\star$ may directly be related to $A(\hat{f}_n) - A(f^\star)$. Applying for example Lemma 2.1 in Zhang (2004), we conclude that with probability at least $1 - 1/n^2$,

$$L(\hat{f}_n) - L^\star \leq 2\sqrt{2C'} \left(\frac{\ln n}{n} \right)^{\frac{\beta}{2\beta+16}}.$$

To understand the significance of this inequality, just recall that what we are after in this article is the supervised classification of (infinite-dimensional) stochastic processes. As enlightened in the proofs, this makes the analysis different from the standard context, where one seeks to learn finite-dimensional quantities. The bridge between the two worlds is crossed via martingale and stochastic calculus arguments. Lastly, it should be noted that the regularity parameter β is assumed to be unknown, so that our results are adaptive as well.

3 Proofs

Throughout this section, if P is a probability measure and f a function, the notation Pf stands for the integral of f with respect to P. By $\mathbb{L}^2(P)$ we mean the space of square integrable real functions with respect to P. Also, for a class $\mathcal{F}$ of functions in $\mathbb{L}^2(P)$ and $\varepsilon > 0$, we denote by $N(\varepsilon, \mathcal{F}, \mathbb{L}^2(P))$ the ε-covering number of $\mathcal{F}$ in $\mathbb{L}^2(P)$, i.e., the minimal number of metric balls of radius ε in $\mathbb{L}^2(P)$ that are needed to cover $\mathcal{F}$ (see, e.g., Definition 2.1.5 in van der Vaart and Wellner, 1996).

3.1 Proof of Theorem 2.1

For any stochastic processes M_1 and M_2, the notation $\mathbb{Q}_{M_2|M_1}$ and $\mathbb{Q}_{M_2}$ respectively mean the distribution under $\mathbb{Q}$ of M_2 given M_1, and the distribution under $\mathbb{Q}$ of M_2.

We start the proof by observing that

$$\mathbb{P}(Y=+1\,|\,X^\tau=x,Z=z)=p_+\frac{\mathrm{d}\mathbb{P}_{X^\tau,Z|Y=+1}}{\mathrm{d}\mathbb{P}_{X^\tau,Z}}(x,z). \tag{3.1}$$

Thus, to prove the theorem, we need to evaluate the above Radon-Nikodym density. To this aim, we introduce the conditional probabilities $\mathbb{P}^\pm=\mathbb{P}(.|Y=\pm 1)$. For any path z of Z, the conditional distributions $\mathbb{P}^+_{X|Z=z}$ and $\mathbb{P}^-_{X|Z=z}$ are those of Poisson processes with intensity $\lambda_+(.,z)$ and $\lambda_-(.,z)$, respectively. Consequently, according to Lemma A.1, the stopped process X^τ satisfies

$$D_+(x,z)\,\mathbb{P}^+_{X^\tau|Z=z}(\mathrm{d}x)=D_-(x,z)\,\mathbb{P}^-_{X^\tau|Z=z}(\mathrm{d}x),$$

where

$$D_\pm(x,z)=\exp\left(-\int_0^{T\wedge\tau}(1-\lambda_\pm(s,z_s))\,\mathrm{d}s-\int_0^{T\wedge\tau}\ln\lambda_\pm(s,z_s)\mathrm{d}x_s\right).$$

Therefore,

$$D_+(x,z)\,\mathbb{P}^+_{X^\tau|Z=z}\otimes\mathbb{P}_Z(\mathrm{d}x,\mathrm{d}z)=D_-(x,z)\,\mathbb{P}^-_{X^\tau|Z=z}\otimes\mathbb{P}_Z(\mathrm{d}x,\mathrm{d}z).$$

But, by independence of Y and Z, one has $\mathbb{P}_Z=\mathbb{P}^+_Z=\mathbb{P}^-_Z$. Thus,

$$\mathbb{P}_{X^\tau,Z|Y=\pm1}(\mathrm{d}x,\mathrm{d}z)=\mathbb{P}^\pm_{X^\tau|Z=z}\otimes\mathbb{P}_Z(\mathrm{d}x,\mathrm{d}z),$$

whence

$$D_+(x,z)\mathbb{P}_{X^\tau,Z|Y=+1}(\mathrm{d}x,\mathrm{d}z)=D_-(x,z)\mathbb{P}_{X^\tau,Z|Y=-1}(\mathrm{d}x,\mathrm{d}z).$$

On the other hand,

$$\mathbb{P}_{X^\tau,Z}(x,z)=p_+\mathbb{P}_{X^\tau,Z|Y=+1}(x,z)+p_-\mathbb{P}_{X^\tau,Z|Y=-1}(x,z),$$

so that

$$\frac{\mathrm{d}\mathbb{P}_{X^\tau,Z|Y=+1}}{\mathrm{d}\mathbb{P}_{X^\tau,Z}}(x,z)=\frac{1}{p_-\frac{D_+(x,z)}{D_-(x,z)}+p_+}.$$

Using identity (3.1), we obtain

$$\mathbb{P}(Y=+1\,|\,X^\tau,Z)=\frac{p_+}{p_-e^{-\xi}+p_+},$$

where

$$\begin{aligned}\xi&=\int_0^{T\wedge\tau}(\lambda_--\lambda_+)(s,Z_s)\mathrm{d}s+\int_0^{T\wedge\tau}\ln\frac{\lambda_+}{\lambda_-}(s,Z_s)\mathrm{d}X_s\\&=\int_0^{T\wedge\tau}(\lambda_--\lambda_+)(s,Z^\tau_s)\mathrm{d}s+\int_0^{T\wedge\tau}\ln\frac{\lambda_+}{\lambda_-}(s,Z^\tau_s)\mathrm{d}X^\tau_s.\end{aligned}$$

Clearly, the random variable ξ is $\sigma(\tau, X^\tau, Z^\tau)$-measurable. Since $\sigma(\tau) \subset \sigma(X^\tau)$, it is also $\sigma(X^\tau, Z^\tau)$-measurable. This observation, combined with the inclusion $\sigma(X^\tau, Z^\tau) \subset \sigma(X^\tau, Z)$, leads to

$$\mathbb{P}(Y = +1 \,|\, X^\tau, Z^\tau) = \mathbb{E}\left(\mathbb{P}(Y = +1 \,|\, X^\tau, Z) \,\middle|\, X^\tau, Z^\tau\right) = \mathbb{P}(Y = +1 \,|\, X^\tau, Z).$$

This shows the desired result.□

3.2 Proof of Theorem 2.2

Theorem 2.2 is mainly a consequence of a general model selection result due to Blanchard et al. (2008), which is recalled in the Appendix for the sake of completeness (Theorem A.1). Throughout the proof, the letter C denotes a generic universal positive constant, whose value may change from line to line. We let $\ell(f)$ be a shorthand notation for the function

$$(x, z, y) \in \mathcal{X} \times \mathcal{Z} \times \{-1, 1\} \mapsto \phi(-yf(x, z)),$$

and let P be the distribution of the prototype triplet (X^τ, Z^τ, Y).

To frame our problem in the vocabulary of Theorem A.1, we consider the family of models $(\mathcal{F}_{B_k})_{k\geq 1}$ and start by verifying that assumptions (i) to (iv) are satisfied. If we define

$$\mathbf{d}^2(f, f') = P\left(\ell(f) - \ell(f')\right)^2,$$

then assumption (i) is immediately satisfied. A minor modification of the proof of Lemma 19 in Blanchard et al. (2003) reveals that, for all integers $B > 0$ and all $f \in \mathcal{F}_B$,

$$P\left(\ell(f) - \ell(f^\star)\right)^2 \leq \left(\phi(UB) + \phi(-UB) + 2 - 2\ln 2\right) P\left(\ell(f) - \ell(f^\star)\right).$$

This shows that assumption (ii) is satisfied with $C_k = 2(\phi(UB_k) + 1 - \ln 2)$. Moreover, it can be easily verified that assumption (iii) holds with $b_k = \phi(UB_k)$.

The rest of the proof is devoted to the verification of assumption (iv). To this aim, for all $B > 0$ and all $f_0 \in \mathcal{F}_B$, we need to bound the expression

$$F_B(r) = \mathbb{E}\sup\left\{\left|(P_n - P)\left(\ell(f) - \ell(f_0)\right)\right| : f \in \mathcal{F}_B, \mathbf{d}^2(f, f_0) \leq r\right\},$$

where

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\left(X_i^{\tau_i}, Z_i^{\tau_i}, Y_i\right)}$$

is the empirical distribution associated to the sample. Let

$$\mathcal{G}_{B,f_0} = \left\{\ell(f) - \ell(f_0) : f \in \mathcal{F}_B\right\}.$$

Then

$$F_B(r) = \mathbb{E}\sup\left\{|(P_n - P)g| : g \in \mathcal{G}_{B,f_0}, Pg^2 \leq r\right\}.$$

Using the symmetrization inequality presented in Theorem 2.1 of Koltchinskii (2011), it is easy to see that

$$F_B(r) \leq 2\,\mathbb{E}\sup\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma_i g\left(X_i^{\tau_i}, Z_i^{\tau_i}, Y_i\right) : g \in \mathcal{G}_{B,f_0}, Pg^2 \leq r\right\}, \qquad (3.2)$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables (that is, $\mathbb{P}(\sigma_i = \pm 1) = 1/2$), independent from the $(X_i^{\tau_i}, Z_i^{\tau_i}, Y_i)$'s. Now, since the functions in $\mathcal{F}_B$ take their values in $[-UB, UB]$, and since ϕ is Lipschitz on this interval with constant $\phi'(UB)$, we have, for all $f, f' \in \mathcal{F}_B$,

$$\sqrt{P_n\left(\ell(f) - \ell(f')\right)^2} \leq \phi'(UB)\sqrt{P_n\left(f - f'\right)^2}.$$

Consequently, for all $\varepsilon > 0$,

$$N\left(2\varepsilon UB\phi'(UB), \mathcal{G}_{B,f_0}, \mathbb{L}^2\left(P_n\right)\right) \leq N\left(2\varepsilon UB, \mathcal{F}_B, \mathbb{L}^2\left(P_n\right)\right).$$

Since $\mathcal{F}_B$ is included in a linear space of dimension at most $2B+1$, Lemma 2.6.15 in van der Vaart and Wellner (1996) indicates that it is a VC-subgraph class of VC-dimension at most $2B+3$. Observing that the function constantly equal to $2UB$ is a measurable envelope for $\mathcal{F}_B$, we conclude from Theorem 9.3 in Kosorok (2008) that, for all $\varepsilon > 0$,

$$N\left(2\varepsilon UB, \mathcal{F}_B, \mathbb{L}^2\left(P_n\right)\right) \leq C\left(2B+3\right)\left(4e\right)^{2B+3}\left(\frac{1}{\varepsilon}\right)^{4(B+1)}.$$

Therefore,

$$N\left(2\varepsilon UB\phi'(UB), \mathcal{G}_{B,f_0}, \mathbb{L}^2\left(P_n\right)\right) \leq C\left(2B+3\right)\left(4e\right)^{2B+3}\left(\frac{1}{\varepsilon}\right)^{4(B+1)}.$$

Now, notice that the constant function equal to $2UB\phi'(UB)$ is a measurable envelope for $\mathcal{G}_{B,f_0}$. Thus, applying Lemma A.2 yields

$$F_B(r) \leq \psi_B(r),$$

where ψ_B is defined for all $r > 0$ by

$$\psi_B(r) = \frac{C\sqrt{r}}{\sqrt{n}} \sqrt{B \ln\left(\frac{A'_B}{\sqrt{r}}\right)} \vee \frac{CBA_B}{n} \ln\left(\frac{A'_B}{\sqrt{r}}\right) \vee \frac{CA_B}{n} \sqrt{B \ln\left(\frac{A'_B}{\sqrt{r}}\right)},$$

with $A_B = UB\phi'(UB)$ and $A'_B = A_B((2B+3)(4e)^{2B+3})^{1/4(B+1)}$. (Notation $t_1 \vee t_2$ means the maximum of t_1 and t_2.)

Attention shows that ψ_B is a sub-root function and assumption (iv) is therefore satisfied. It is routine to verify that the solution $r_k^\star$ of $\psi_{B_k}(r) = r/C_k$ satisfies, for all $k \geq 1$ and all $n \geq 1$,

$$r_k^\star \leq C \left(A_{B_k}^2 B_k C_k^2 + \sqrt{A'_{B_k}}\right) \frac{\ln n}{n}.$$

Furthermore, observing that the function $B \mapsto ((2B+3)(4e)^{2B+3})^{1/4(B+1)}$ is bounded from above, we obtain

$$r_k^\star \leq C \left(A_{B_k}^2 B_k C_k^2 + \sqrt{A_{B_k}}\right) \frac{\ln n}{n}.$$

Hence, taking $x_k = \alpha \ln \lambda_k$ and $K = 11/5$ in Theorem A.1, and letting

$$R_k = A_{B_k}^2 B_k C_k + \frac{\sqrt{A_{B_k}}}{C_k},$$

we conclude that there exists a universal constant $C > 0$ such that, if the penalty $\mathrm{pen} : \mathbb{N}^\star \to \mathbb{R}_+$ satisfies

$$\mathrm{pen}(k) \geq C \left\{ R_k \frac{\ln n}{n} + \frac{C_k \left(\alpha \ln B_k + \delta + \ln 2\right)}{n} \right\}$$

for some $\delta > 0$, then, with probability at least $1 - e^{-\delta}$,

$$A(\hat{f}_n) - A(f^\star) \leq 2 \inf_{k \geq 1} \left\{ \inf_{f \in \mathcal{F}_{B_k}} \left(A(f) - A(f^\star)\right) + \mathrm{pen}(k) \right\}.$$

This completes the proof. □

3.3 Proof of Proposition 2.1

Proof of Proposition 2.1 relies on the following intermediary lemma, which is proved in the next subsection.

Lemma 3.1 *Assume that Assumption* **A** *holds. Then, for all positive integers $B \geq 1$,*

$$\begin{aligned}\inf_{f\in\mathcal{F}_B}\left(A(f)-A(f^\star)\right) \leq\ & 2D\min\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-(\lambda_- -\lambda_+)\right\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}\\ &+4D\min\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-\ln\frac{\lambda_+}{\lambda_-}\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}\\ &+2\min_{|x|\leq B}\left|x-\ln\frac{p_+}{p_-}\right|,\end{aligned}$$

where the first two minima are taken over all $\alpha=(\alpha_1,\ldots,\alpha_B)\in\mathbb{R}^B$ with $\sum_{j=1}^{B}|\alpha_j|\leq B$.

Proof of Proposition 2.1 – It can be easily verified that, for all $f\in\mathbb{L}^2(\mathrm{d}s\otimes\mu)$, one has

$$\|f\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}\leq\frac{\|f\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}}{\sqrt{T\mu\left([0,1]^d\right)}}.$$

Consequently, for all $B\geq 1$,

$$\begin{aligned}&\min\left\{\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-(\lambda_- -\lambda_+)\right\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}:\sum_{j=1}^{B}|\alpha_j|\leq B\right\}\\ &\quad\leq\min\left\{\frac{1}{\sqrt{T\mu\left([0,1]^d\right)}}\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-(\lambda_- -\lambda_+)\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}:\sum_{j=1}^{B}|\alpha_j|\leq B\right\}\\ &\quad\leq\min\left\{\frac{1}{\sqrt{T\mu\left([0,1]^d\right)}}\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-(\lambda_- -\lambda_+)\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}:\sum_{j=1}^{B}\alpha_j^2\leq B\right\}.\end{aligned}\tag{3.3}$$

Since $\lambda_- -\lambda_+\in\mathcal{W}(\beta,M)$ and $B\geq M^2$, we have

$$\sum_{j=1}^{B}a_j^{\star 2}\leq\sum_{j=1}^{\infty}j^{2\beta}a_j^{\star 2}\leq M^2\leq B.\tag{3.4}$$

Thus, combining (3.3) and (3.4) yields, for $B \geq M^2$,

$$\begin{aligned}
&\min\left\{\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-(\lambda_--\lambda_+)\right\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}:\sum_{j=1}^{B}|\alpha_j|\leq B\right\}\\
&\quad\leq\frac{1}{\sqrt{T\mu\left([0,1]^d\right)}}\left\|\sum_{j=1}^{B}a_j^\star\varphi_j-(\lambda_--\lambda_+)\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}\\
&\quad=\frac{1}{\sqrt{T\mu\left([0,1]^d\right)}}\left\|\sum_{j=B+1}^{\infty}a_j^\star\varphi_j\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}. \qquad (3.5)
\end{aligned}$$

It follows from the properties of an orthonormal basis and the definition of $\mathcal{W}(\beta,M)$ that

$$\begin{aligned}
\left\|\sum_{j=B+1}^{\infty}a_j^\star\varphi_j\right\|^2_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}&=\sum_{j=B+1}^{\infty}a_j^{\star2}\\
&\leq\sqrt{\sum_{j=B+1}^{\infty}j^{2\beta}a_j^{\star2}}\sqrt{\sum_{j=B+1}^{\infty}\frac{a_j^{\star2}}{j^{2\beta}}}\\
&\leq M\sqrt{\sum_{j=B+1}^{\infty}\frac{a_j^{\star2}}{j^{2\beta}}}\\
&\leq\frac{M\|a^\star\|_2}{B^\beta}. \qquad (3.6)
\end{aligned}$$

Inequalities (3.5) and (3.6) show that, for all $B\geq M^2$,

$$\min\left\{\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-(\lambda_--\lambda_+)\right\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}:\sum_{j=1}^{B}|\alpha_j|\leq B\right\}\leq\sqrt{\frac{M\|a^\star\|_2}{T\mu\left([0,1]^d\right)B^\beta}}.$$

Similarly, it may be proved that, for all $B\geq M^2$,

$$\min\left\{\left\|\sum_{j=1}^{B}\alpha_j\varphi_j-\ln\frac{\lambda_+}{\lambda_-}\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}:\sum_{j=1}^{B}|\alpha_j|\leq B\right\}\leq\sqrt{\frac{M\|b^\star\|_2}{B^\beta}}.$$

Applying Lemma 3.1 we conclude that, whenever $B\geq\max(M^2,\ln\frac{p_+}{p_-})$,

$$\inf_{f\in\mathcal{F}_B}\left(A(f)-A(f^\star)\right)\leq 2D\sqrt{\frac{M\|a^\star\|_2}{T\mu\left([0,1]^d\right)B^\beta}}+4D\sqrt{\frac{M\|b^\star\|_2}{B^\beta}}.\ \square$$

3.4 Proof of Lemma 3.1

We start with a technical lemma.

Lemma 3.2 *Let $\phi(t) = \ln_2(1 + e^t)$ be the logit loss. Then*

$$\arg\min_f \mathbb{E}\phi\left(-Yf(X^\tau, Z^\tau) \,|\, X^\tau, Z^\tau\right) = \xi - \ln\frac{p_-}{p_+},$$

where the minimum is taken over all measurable functions $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$.

Proof – According to the results of Section 2.2 in Bartlett et al. (2006), one has

$$\arg\min_f \mathbb{E}\phi\left(-Yf(X^\tau, Z^\tau) \,|\, X^\tau, Z^\tau\right) = \alpha^\star\left(\eta(X^\tau, Z^\tau)\right),$$

where, for all $0 \leq \eta \leq 1$,

$$\alpha^\star(\eta) = \arg\min_{\alpha\in\mathbb{R}} \left(\eta\phi(-\alpha) + (1-\eta)\phi(\alpha)\right).$$

With our choice for ϕ, it is straightforward to check that, for all $0 \leq \eta < 1$,

$$\alpha^\star(\eta) = \ln\left(\frac{\eta}{1-\eta}\right).$$

Since, by assumption, $p_- > 0$, we have

$$\eta(X^\tau, Z^\tau) = \frac{p_+}{p_- e^{-\xi} + p_+} < 1.$$

Thus

$$\alpha^\star\left(\eta(X^\tau, Z^\tau)\right) = \xi - \ln\frac{p_-}{p_+},$$

which is the desired result. □

Proof of Lemma 3.1 – Let $B > 0$ be fixed. Let $a_1, \ldots, a_B$ and $b_1, \ldots, b_B$ be real numbers such that

$$\left\|\sum_{j=1}^B a_j\varphi_j - (\lambda_- - \lambda_+)\right\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)} = \min\left\|\sum_{j=1}^B \alpha_j\varphi_j - (\lambda_- - \lambda_+)\right\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}$$

and

$$\left\|\sum_{j=1}^B b_j\varphi_j - \ln\frac{\lambda_+}{\lambda_-}\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)} = \min\left\|\sum_{j=1}^B \alpha_j\varphi_j - \ln\frac{\lambda_+}{\lambda_-}\right\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)},$$

where, in each case, the minimum is taken over all $\alpha = (\alpha_1, \ldots, \alpha_B) \in \mathbb{R}^B$ with $\sum_{j=1}^B |\alpha_j| \leq B$. Let also $c \in \mathbb{R}$ be such that

$$\left| c - \ln \frac{p_+}{p_-} \right| = \min_{|x| \leq B} \left| x - \ln \frac{p_+}{p_-} \right|.$$

Introduce f_B, the function in $\mathcal{F}_B$ defined by

$$\begin{aligned} f_B &= \sum_{j=1}^B [a_j \Phi_j + b_j \Psi_j] + c \\ &= \int_0^{T \wedge \tau} \sum_{j=1}^B a_j \varphi_j(s, Z_s) \mathrm{d}s + \int_0^{T \wedge \tau} \sum_{j=1}^B b_j \varphi_j(s, Z_s) \mathrm{d}X_s + c. \end{aligned}$$

Clearly,

$$\inf_{f \in \mathcal{F}_B} (A(f) - A(f^\star)) \leq A(f_B) - A(f^\star).$$

Since ϕ is Lipschitz with constant $\phi'(UB) = (\ln 2(1 + e^{-UB}))^{-1} \leq 2$ on the interval $[-UB, UB]$, we have

$$|A(f_B) - A(f^\star)| \leq 2\mathbb{E} \left| f_B(X^\tau, Z^\tau) - f^\star(X^\tau, Z^\tau) \right|. \tag{3.7}$$

But, by Lemma 3.2,

$$f^\star(X^\tau, Z^\tau) = \int_0^{T \wedge \tau} (\lambda_- - \lambda_+)(s, Z_s) \mathrm{d}s + \int_0^{T \wedge \tau} \ln \frac{\lambda_+}{\lambda_-}(s, Z_s) \mathrm{d}X_s + \ln \frac{p_+}{p_-}.$$

Thus, letting,

$$\vartheta_1 = \sum_{j=1}^B a_j \varphi_j - (\lambda_- - \lambda_+) \quad \text{and} \quad \vartheta_2 = \sum_{j=1}^B b_j \varphi_j - \ln \frac{\lambda_+}{\lambda_-},$$

it follows

$$\begin{aligned} \mathbb{E} \left| f_B(X^\tau, Z^\tau) - f^\star(X^\tau, Z^\tau) \right| \leq\ & \mathbb{E} \left| \int_0^{T \wedge \tau} \vartheta_1(s, Z_s) \mathrm{d}s \right| \\ & + \mathbb{E} \left| \int_0^{T \wedge \tau} \vartheta_2(s, Z_s) \mathrm{d}X_s \right| \\ & + \left| c - \ln \frac{p_+}{p_-} \right|. \end{aligned} \tag{3.8}$$

Since the distribution $\mathbb{P}_{Z_s}$ of Z_s has a density h_s with respect to μ, and since this density is uniformly bounded by D, we obtain

$$\begin{aligned}\mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_1(s,Z_s)\mathrm{d}s\right| &\leq \int_0^T\int_{[0,1]^d}|\vartheta_1(s,z)|\,\mathbb{P}_{Z_s}(\mathrm{d}z)\mathrm{d}s\\ &= \int_0^T\int_{[0,1]^d}|\vartheta_1(s,z)|\,h_s(z)\mu(\mathrm{d}z)\mathrm{d}s\\ &\leq D\|\vartheta_1\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}. \end{aligned} \tag{3.9}$$

With a slight abuse of notation, set $\lambda_Y = \lambda_\pm$, depending on whether $Y = \pm 1$, and

$$\Lambda_{Y,Z}(t) = \int_0^t \lambda_Y(s,Z_s)\mathrm{d}s, \quad t\in[0,T].$$

With this notation,

$$\begin{aligned}\mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\mathrm{d}X_s\right| &\leq \mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\mathrm{d}\left(X_s-\Lambda_{Y,Z}(s)\right)\right|\\ &\quad+\mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\mathrm{d}\Lambda_{Y,Z}(s)\right|\\ &= \mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\mathrm{d}\left(X_s-\Lambda_{Y,Z}(s)\right)\right|\\ &\quad+\mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\lambda_Y(s)\mathrm{d}s\right|. \end{aligned} \tag{3.10}$$

Since $X-\Lambda_{Y,Z}$ is a martingale conditionally to Y and Z, the Ito isometry (see Theorem I.4.40 in Jacod and Shiryaev, 2003) yields

$$\begin{aligned}&\mathbb{E}\left[\left(\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\mathrm{d}\left(X_s-\Lambda_{Y,Z}(s)\right)\right)^2\,\middle|\,Y,Z\right]\\ &\quad=\mathbb{E}\left[\int_0^{T\wedge\tau}\vartheta_2^2(s,Z_s)\mathrm{d}\langle X_s-\Lambda_{Y,Z}(s)\rangle\,\middle|\,Y,Z\right], \end{aligned} \tag{3.11}$$

where $\langle M\rangle$ stands for the predictable compensator of the martingale M. Observing that X is a Poisson process with intensity $s\mapsto\lambda_Y(s,Z_s)$ conditionally to Y and Z, we deduce that $\langle X-\Lambda_{Y,Z}\rangle = \langle X\rangle = \Lambda_{Y,Z}$ conditionally to Y and Z. As a result,

$$\begin{aligned}&\mathbb{E}\left[\int_0^{T\wedge\tau}\vartheta_2^2(s,Z_s)\mathrm{d}\langle X_s-\Lambda_{Y,Z}(s)\rangle\,\middle|\,Y,Z\right]\\ &\quad=\mathbb{E}\left[\int_0^{T\wedge\tau}\vartheta_2^2(s,Z_s)\lambda_Y(s,Z_s)\mathrm{d}s\,\middle|\,Y,Z\right]. \end{aligned} \tag{3.12}$$

Combining (3.10)-(3.12) we deduce that

$$\mathbb{E}\left|\int_0^{T\wedge\tau}\vartheta_2(s,Z_s)\mathrm{d}X_s\right| \leq 2D\|\vartheta_2\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}. \tag{3.13}$$

Putting together identities (3.7), (3.8), (3.9) and (3.13) yields

$$\begin{aligned}&\inf_{f\in\mathcal{F}_B}\left(A(f)-A(f^\star)\right)\\ &\quad\leq 2D\|\vartheta_1\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)}+4D\|\vartheta_2\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}+2\min_{|x|\leq B}\left|x-\ln\frac{p_+}{p_-}\right|,\end{aligned}$$

which concludes the proof by definition of ϑ_1 and ϑ_2. □

A Appendix

A.1 A general theorem for model selection

The objective of this section is to recall a general model selection result due to Blanchard et al. (2008).

Let $\mathcal{X}$ be a measurable space and let $\ell:\mathbb{R}\times\{-1,1\}\to\mathbb{R}$ be a loss function. Given a function $g:\mathcal{X}\to\mathbb{R}$, we let $\ell(g)$ be a shorthand notation for the function $(x,y)\in\mathbb{R}\times\{-1,1\}\mapsto\ell(g(x),y)$. Let P be a probability distribution on $\mathcal{X}\times\{-1,1\}$ and let $\mathfrak{G}$ be a set of extended-real valued functions on $\mathcal{X}$ such that, for all $g\in\mathfrak{G}$, $\ell(g)\in\mathbb{L}^2(P)$. The target function $g^\star$ is defined as

$$g^\star\in\arg\min_{g\in\mathfrak{G}}P\ell(g).$$

Let $(\mathcal{G}_k)_{k\geq1}$ be a countable family of models such that, for all $k\geq1$, $\mathcal{G}_k\subset\mathfrak{G}$. For each $k\geq1$, we define the empirical risk minimizer $\hat{g}_k$ as

$$\hat{g}_k\in\arg\min_{g\in\mathcal{G}_k}P_n\ell(g).$$

If pen denotes a real-valued function on $\mathbb{N}^\star$, we let the penalized empirical risk minimizer $\hat{g}$ be defined by $\hat{g}_{\hat{k}}$, where

$$\hat{k}\in\arg\min_{k\geq1}\left[P_n\ell(\hat{g}_k)+\mathrm{pen}(k)\right].$$

Recall that a function $\mathbf{d}:\mathfrak{G}\times\mathfrak{G}\to\mathbb{R}_+$ is a pseudo-distance if (i) $\mathbf{d}(g,g)=0$, (ii) $\mathbf{d}(g,g')=\mathbf{d}(g',g)$, and (iii) $\mathbf{d}(g,g')\leq\mathbf{d}(g,g'')+\mathbf{d}(g'',g')$ for all g, g', g'' in $\mathfrak{G}$. Also, a function $\psi:\mathbb{R}_+\to\mathbb{R}_+$ is said to be a sub-root function if (i) it is nondecreasing and (ii) the function $r\in\mathbb{R}_+\mapsto\psi(r)/\sqrt{r}$ is nonincreasing.

Theorem A.1 (Blanchard et al., 2008) *Assume that there exist a pseudo-distance* $\mathbf{d}$ *on* $\mathfrak{G}$*, a sequence of sub-root functions* $(\psi_k)_{k\geq 1}$*, and two non-decreasing sequences* $(b_k)_{k\geq 1}$ *and* $(C_k)_{k\geq 1}$ *of real numbers such that*

(i) $\forall g, g' \in \mathfrak{G} : P(\ell(g) - \ell(g'))^2 \leq \mathbf{d}^2(g, g')$*;*

(ii) $\forall k \geq 1$*,* $\forall g \in \mathcal{G}_k : \mathbf{d}^2(g, g^\star) \leq C_k P(\ell(g) - \ell(g^\star))$*;*

(iii) $\forall k \geq 1$*,* $\forall g \in \mathcal{G}_k$*,* $\forall (x, y) \in \mathcal{X} \times \{-1, 1\} : |\ell(g(x), y)| \leq b_k$*;*

and, if $r_k^\star$ *denotes the solution of* $\psi_k(r) = r/C_k$*,*

(iv) $\forall k \geq 1$*,* $\forall g_0 \in \mathcal{G}_k$*,* $\forall r \geq r_k^\star$ *:*

$$\mathbb{E} \sup \left\{ |(P_n - P)(\ell(g) - \ell(g_0))| : g \in \mathcal{G}_k, \mathbf{d}^2(g, g_0) \leq r \right\} \leq \psi_k(r).$$

Let $(x_k)_{k\geq 1}$ *be a nonincreasing sequence such that* $\sum_{k\geq 1} e^{-x_k} \leq 1$*. Let* $\delta > 0$ *and* $K > 1$ *be two fixed real numbers. If* $\mathrm{pen}(k)$ *denotes a penalty term satisfying*

$$\forall k \geq 1, \quad \mathrm{pen}(k) \geq 250K \frac{r_k^\star}{C_k} + \frac{(65KC_k + 56b_k)(x_k + \delta + \ln 2)}{3n},$$

then, with probability at least $1 - e^{-\delta}$*, one has*

$$P(\ell(\hat{g}) - \ell(g^\star)) \leq \frac{K + \frac{1}{5}}{K - 1} \inf_{k\geq 1} \left\{ \inf_{g \in \mathcal{G}_k} P(\ell(g) - \ell(g^\star)) + 2\mathrm{pen}(k) \right\}.$$

A.2 Expected supremum of Rademacher processes

Let S be a measurable space and let P be a probability measure on S. Let $\mathcal{G}$ be a class of functions $g : S \to \mathbb{R}$. The Rademacher process $(R_n(g))_{g\in\mathcal{G}}$ associated with P and indexed by $\mathcal{G}$ is defined by

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(Z_i),$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables, and $Z_1, \ldots, Z_n$ is a sequence of i.i.d. random variables, with distribution P and independent of the σ_i's.

We recall in this subsection a bound for the supremum of the Rademacher process defined by

$$\|R_n\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |R_n(g)|,$$

which follows from the results of Giné and Koltchinskii (2006). Let G be a measurable envelope for $\mathcal{G}$, i.e., a measurable function $G : S \to \mathbb{R}_+$ such that

$$\sup_{x \in S} |g(x)| \leq G(x).$$

Define $\|G\| = \sqrt{PG^2}$ and $\|G\|_n = \sqrt{P_n G^2}$, where $P_n = n^{-1} \sum_{i=1}^n \delta_{Z_i}$ stands for the empirical measure associated to $Z_1, \ldots, Z_n$. Finally, let $\sigma^2 > 0$ be a real number satisfying

$$\sup_{g \in \mathcal{G}} Pg^2 \leq \sigma^2 \leq \|G\|^2.$$

Theorem A.2 (Giné and Koltchinskii, 2006) *Assume that the functions in $\mathcal{G}$ are uniformly bounded by a constant $U > 0$. Assume, in addition, that there exist two constants C and $V > 0$ such that, for all $n \geq 1$ and all $0 < \epsilon \leq 2$,*

$$N\left(\varepsilon \|G\|_n, \mathcal{G}, \mathbb{L}^2\left(P_n\right)\right) \leq \left(\frac{C}{\varepsilon}\right)^V.$$

Then, for all $n \geq 1$,

$$\mathbb{E}\|R_n\|_{\mathcal{G}} \leq \frac{c\sigma}{\sqrt{n}} \sqrt{V \ln\left(\frac{c'\|G\|}{\sigma}\right)} \vee \frac{8c^2 UV}{n} \ln\left(\frac{c'\|G\|}{\sigma}\right) \vee \frac{cU}{9n} \sqrt{V \ln\left(\frac{c'\|G\|}{\sigma}\right)},$$

where $c = 432$ and $c' = 2e \vee C$.

A.3 Some stochastic calculus results

Lemma A.1 *Let μ (resp., ν) be the distribution of a Poisson process on $[0, T]$ with intensity $\lambda : [0, T] \to \mathbb{R}_+^\star$ (resp., with intensity 1) stopped after its u-th jump. Then, μ and ν are equivalent. Moreover,*

$$\nu(\mathrm{d}x) = \exp\left(-\int_0^{T \wedge \tau(x)} (1 - \lambda(s))\, \mathrm{d}s - \int_0^{T \wedge \tau(x)} \ln \lambda(s) \mathrm{d}x_s\right) \mu(\mathrm{d}x),$$

where, for all $x \in \mathcal{X}$, $\tau(x) = \inf\{t \in [0, T] : x_t = u\}$.

Proof. Consider the canonical Poisson process $N = (N_t)_{t \in [0,T]}$ with intensity λ on the filtered space $(\mathcal{X}, (\mathcal{A}_t)_{t \in [0,T]}, \mathbb{P})$, where $\mathcal{A}_t = \sigma(N_s : s \in [0, t])$, and let, for all $t \in [0, T]$,

$$\Lambda(t) = \int_0^t \lambda(s) \mathrm{d}s \quad \text{and} \quad h(t) = \frac{1}{\lambda(t)} - 1.$$

Recall that the process $M = (M_t)_{t\in[0,T]}$ defined by $M_t = N_t - \Lambda(t)$ is a martingale. The Doléans-Dade exponential $\mathcal{E} = (\mathcal{E}_t)_{t\in[0,T]}$ of the martingale $h.M$ (see, e.g., Theorem I.4.61 in Jacod and Shiryaev, 2003) is defined for all $t \in [0,T]$ by

$$\begin{aligned}\mathcal{E}_t &= e^{h.M_t}\prod_{s\leq t}(1+\Delta h.M_s)e^{-\Delta h.M_s}\\ &= \exp\left(-\int_0^t h(s)\lambda(s)\mathrm{d}s + \int_0^t \ln\left(1+h(s)\right)\mathrm{d}N_s\right)\\ &= \exp\left(-\int_0^t (1-\lambda(s))\,\mathrm{d}s - \int_0^t \ln\lambda(s)\mathrm{d}N_s\right), \qquad \text{(A.1)}\end{aligned}$$

where $\Delta h.M_s = h.M_s - h.M_{s^-} = h.N_s - h.N_{s^-}$. Equivalently, $\mathcal{E}$ is the solution to the stochastic equation

$$\mathcal{E} = 1 + \mathcal{E}^-.(h.M) = 1 + (\mathcal{E}^- h).M,$$

where $\mathcal{E}^-$ stands for the process defined by $\mathcal{E}_t^- = \mathcal{E}_{t^-}$. In particular, $\mathcal{E}$ is a martingale. Observe also, since N is a counting process, that the quadratic covariation between M and $\mathcal{E}$ is

$$[M,\mathcal{E}] = (\mathcal{E}^- h).[N,N] = (\mathcal{E}^- h).N.$$

Consequently,

$$[M,\mathcal{E}] - (\mathcal{E}^- h).\Lambda = (\mathcal{E}^- h).M$$

is a martingale. Since $(\mathcal{E}^- h).\Lambda$ is a continuous and adapted process, it is a predictable process and the predictable compensator of $[M,\mathcal{E}]$ takes the form $\langle M,\mathcal{E}\rangle = (\mathcal{E}^- h).\Lambda$. Now let $\mathbb{Q}$ be the measure defined by

$$\mathrm{d}\mathbb{Q} = \mathcal{E}_T \mathrm{d}\mathbb{P}.$$

Since the process $\mathcal{E}$ is a martingale, $\mathbb{Q}$ is a probability and in addition, for all $t \in [0,T]$,

$$\mathrm{d}\mathbb{Q}_t = \mathcal{E}_t \mathrm{d}\mathbb{P}_t, \qquad \text{(A.2)}$$

where $\mathbb{Q}_t$ and $\mathbb{P}_t$ are the respective restrictions of $\mathbb{Q}$ and $\mathbb{P}$ to $\mathcal{A}_t$. Thus, according to the Girsanov theorem (see, e.g., Theorem III.3.11 in Jacod and Shiryaev, 2003), the stochastic process $M - (\mathcal{E}^-)^{-1}.\langle M,\mathcal{E}\rangle$ is a $\mathbb{Q}$-martingale. But, for all $t \in [0,T]$,

$$\begin{aligned}M_t - (\mathcal{E}^-)^{-1}.\langle M,\mathcal{E}\rangle_t &= N_t - \Lambda(t) - h.\Lambda(t)\\ &= N_t - (1+h).\Lambda(t)\\ &= N_t - t.\end{aligned}$$

Thus, the counting process N is such that $(N_t - t)_{t\in[0,T]}$ is a $\mathbb{Q}$-martingale. By the Watanabe theorem (e.g., Theorem IV.4.5 in Jacod and Shiryaev, 2003), this implies that the distribution of N under $\mathbb{Q}$ is that of a Poisson process with unit intensity. So, $\nu = \mathbb{Q}_{T\wedge\tau}$, where $\mathbb{Q}_{T\wedge\tau}$ is the restriction of $\mathbb{Q}$ to the stopped σ-field $\mathcal{A}_{T\wedge\tau}$. Moreover, by Theorem III.3.4 in Jacod and Shiryaev (2003) and identity (A.2), we have

$$\mathrm{d}\mathbb{Q}_{T\wedge\tau} = \mathcal{E}_{T\wedge\tau}\mathrm{d}\mathbb{P}_{T\wedge\tau},$$

where the definition of $\mathbb{P}_{T\wedge\tau}$ is clear. Since $\mu = \mathbb{P}_{T\wedge\tau}$, the result is a consequence of identity (A.1). □

References

P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting methods. *Journal of Machine Learning Research*, 4: 861–894, 2003.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36:489–531, 2008.

P.R. Bouzas, M.J. Valderrama, A.M. Aguilera, and N.R. Ruiz-Fuentes. Modelling the mean of a doubly stochastically Poisson process by functional data analysis. *Computational Statistics & Data Analysis*, 50:2655–2667, 2006.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

B. Cadre. *Supervised classification of diffusion paths.* Preprint, Ecole Normale Supérieure de Cachan, Antenne de Bretagne, Bruz, 2012.

E.J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2005.

D.R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B*, 17:129–164, 1955.

D.R. Cox. Regression modes and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

D.R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

C. Denis. *Classification in postural style based on stochastic process modeling.* hal-00653316, 2012.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

R. Durrett. *Essential of Stochastic Processes.* Springer, New York, 2010.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34: 1143–1216, 2006.

G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes. Third Edition.* Oxford University Press, Oxford, 2001.

J. Illian, E. Benson, J. Crawford, and H. Staines. Principal component analysis for spatial point processes - assessing the appropriateness of the approach in an ecological context. *Lecture Notes in Statistics*, 185:135–150, 2006.

J. Jacod and A.N. Shiryaev. *Limits Theorems for Stochastic Processes. Second Edition.* Springer, New York, 2003.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Springer, Berlin, 2011.

M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York, 2008.

M. Krumin and S. Shoham. Generation of spike trains with controlled auto- and cross-correlation functions. *Neural Computation*, 21:1642–1664, 2009.

D. Lando. On Cox processes and credit risky securities. *Review of Derivatives Research*, 2:99–120, 1998.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods (with discussion). *The Annals of Statistics*, 32:30–55, 2004.

P. Massart and C. Meynet. An ℓ_1-oracle inequality for the Lasso. *Electronic Journal of Statistics*, 5:669–687, 2011.

F. O'Sullivan. Nonparametric estimation in the Cox model. *The Annals of Statistics*, 21:124–145, 1993.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis. Second Edition.* Springer, New York, 2005.

D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion. Third Edition.* Springer, New York, 2005.

R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5: 197–227, 1990.

W. Shuang, H.-G. Müller, and Z. Zhang. Functional data analysis for point processes with rare events. *Statistica Sinica*, 23:1–23, 2013.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288, 1996.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York, 1996.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). *The Annals of Statistics*, 32:56–85, 2004.

B. Zhu, P.X.-K. Song, and J.M.G. Taylor. Stochastic functional data analysis: A diffusion model-based approach. *Biometrics*, 67:1295–1304, 2011.